# Semantic Interoperability

*Think piece for the Science Forum 2009 Workshop on*
*"ICTs transforming agricultural science, research and technology generation"*

*Johannes Keizer (FAO), Valeria Pesce (GFAR)*

## Introduction

This think piece will center on the concept of interoperability, an issue on which our team in FAO and GFAR has been working together for many years.

The real issue nowadays is not **technological interoperability**. At the moment there is a very high degree, albeit still potential, of technological interoperability on the web that has been achieved through some agreed protocols and technologies. There are still occasional hiccups (browser compatibility for example), but overall the World Wide Web has produced a technology layer on top of different hardware and software that lays the basis for communication.

Instead, this think piece will stress the importance of and the issues related to **semantic interoperability**, the possibility that information distributed over the net can be processed by programs that understand the meaning of the bits and bytes in which the information items are encoded without the need for constant human interpretation (and translation). Better than through a lengthy explanation, semantic interoperability can be illustrated by a video on Web 3.0 (http://www.serviceweb30.eu/cms/index.php/service-web-3-0-the-future-internet) that has been produced for the European Commission and explains the challenge very nicely.

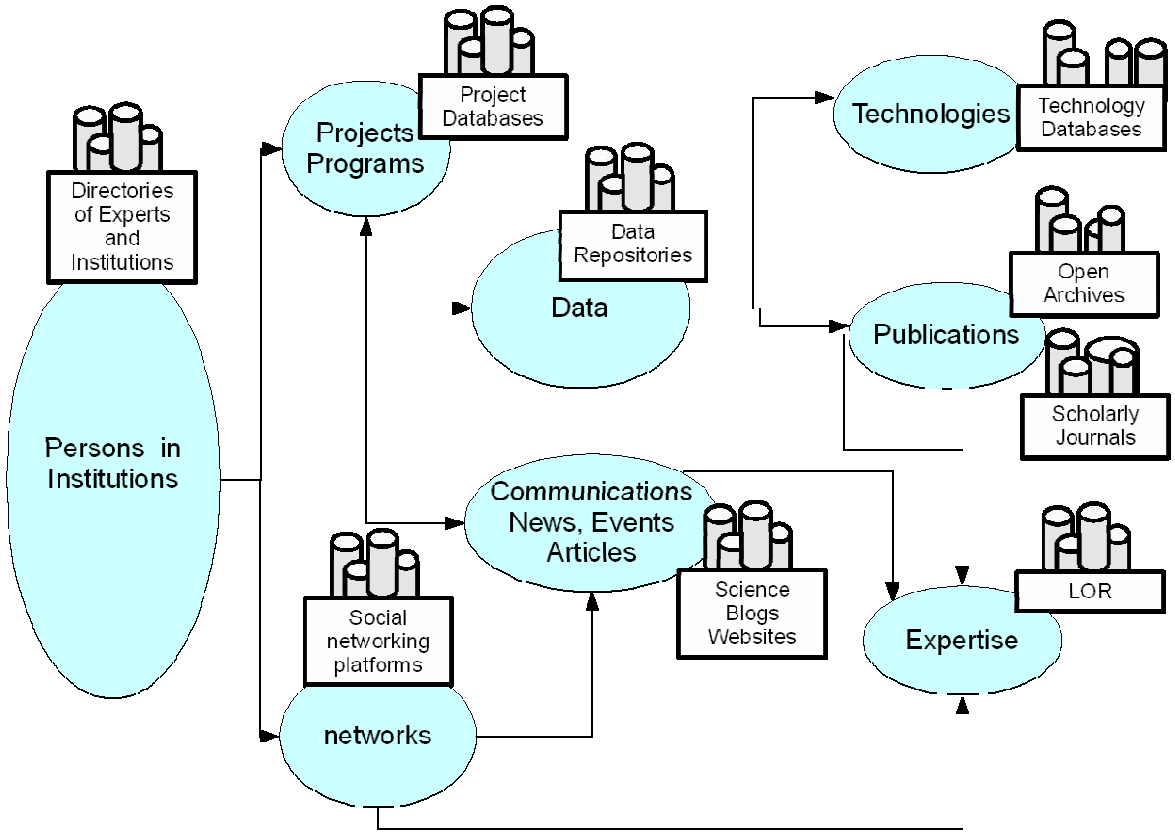
We are quite unsure if there will ever be something like a global sharing of resources. There are problems of attitude ("Why should I share with everyone?", what motivates researchers to share and what doesn't?) and problems of practicality (will we ever be able to agree on global standards and vocabularies to describe scientific knowledge, persons, processes, institutions...?). A good example is the Open Archive Initiative (OAI): this has most probably been the biggest interoperability effort in the area of science, research and innovation in the last 10 years, establishing a new paradigm for sharing ("Open Access"), establishing also a protocol for exchange (OAI PMH) and a semantic metadata exchange standard (simple DC). Nevertheless: a) institutional repositories have not been an overwhelming success until now (thematic repositories have done much better); b) existing harvesting services like OAIster suffer from the semantic scarcity of the Simple DC metadata.

Sharing and interoperability will probably develop in patches, around networks of people who collaborate. The intensity of sharing and the degree of interoperability will vary depending on trust and necessity.

Such networks (communities of practice) come up because of common interests and goals. Under this respect, the most important event in the area of agricultural science and technology has been the establishment in 2008 of the CIARD (Coherence in Information for Agricultural Research for Development) initiative. With CGIAR, GFAR, FAO, and IAALD as the main international sponsors of the initiative, CIARD aims to bring together all actors in the area of agricultural research for development and to persuade them to become partners in making public domain agricultural research

information truly Available, Accessible and Applicable. In this context, the issue of technological and semantic interoperability is being tackled by the CIARD Content Management Task Force (CMTF).

At their last meeting, the CMTF elaborated the following "ecosystem" for the work within the CIARD partner network.




This diagram describes people, processes and technology involved in the production of agricultural innovation knowledge. In this "think piece", some key elements are highlighted that make this ecosystem work.

## Semantic interoperability: vocabularies, tools, capacities.

a) **Ontologies and Vocabularies** as agreed “world views” for interoperability

Unfortunately the word "ontology" has been borrowed from philosophy to describe the necessity of common “worldviews" if someone wants to collaborate. We say unfortunately, because in Philosophy an ontological point of view is in collision with a pragmatic point of view. We need to be very pragmatic if we speak about "ontologies" in Information Science.

If we want to exchange bibliographical data we need to agree on what a "title" is, if we want to do common research on plant mutations, we need to agree about the meaning of a "breakpoint". The semantic web is all about “teaching” machines these agreements and making them able to use them. After such an implementation a search for "title" will return only the titles of publications and not the

titles of persons. If a "breakpoint" is defined as a chromosome alteration that leads to certain mutations, a machine will be able to identify such mutations from a corpus of text.

Being pragmatic also means to understand that for many practical purposes today the agreement on specific vocabularies and very light-weight ontologies would suffice. A classic example is the nuisance caused by giving one's personal information 2-3 times in a week to some application on the web. The problem is, one cannot simply write "johannes" because the system would not know that this is johannes, who works at FAO, lives in Berlin and has that specific credit card number. If we had a directory of persons, in which this information was presented as an Ontology (johannes "has surname", "owns credit card", "lives in", "has telephone", with attributes still defined in another vocabularies), the situation would be much better. The "openID" project partially addresses this issue.

The CIARD CMTF addresses this issue through projects on Application Profiles (AP) and global directories. An example is the Ag-Org Application Profile and foreseen AgriOrg directory. An AP defining a standard description of an organization was developed: using this standard description any organization can create an RDF file and save it under a URI that will be registered in the AgriOrg directory. This means that any service can harvest and reuse this information without new gathering of data and negotiation of meaning. Of course, directories like this will start in the context of a network with similar interests and goals, for instance in the context of the CIARD partnership. Another example of this use case for APs, distributed architectures and directories is the use made of the Ag-Event AP in AgriFeeds. Partners who are using the Ag-Event AP are automatically contributing to a global ARD calendar of events.

Considerable work has been done on textual information objects, which is obviously only a small part of the diagram presented in the introduction of this think piece. Also some work on vocabularies has been done regarding organizations, projects, news and events, which was relatively easy. Working on ontologies for actual research data is a longer process and it is scientific work. For genetic and preservation data the gene and crop ontologies emerged in the last years. The CGIAR now has a major initiative for developing crop ontologies. There are important developments in the GIS area, and there is a slowly but steadily growing basis for semantic interoperability.

On this purpose, it is important to underline again that a pragmatic approach is needed. Fully-fledged ontologies are complex and a lot of work to build. A network of open archives using the AGRIS Application Profile can be sufficient for exchanging publications, a "folksonomy" can be of use for sharing news. However, a rice gene ontology will be necessary to organize a common research project on the introduction of genes into rice to increase the protein content.

But "light-weight" and "full-fledged" approaches are not necessarily alternative to each other. On the contrary, interoperability can gain a lot from the fact that nowadays the work on ontologies and controlled vocabularies on the one hand and the work on folksonomies and natural language processing (NLP) on the other are moving towards each other. Thanks to the mapping between ontologies and to the fact that NLP can make use of ontologies for its reasoning, in the future a user tagging a resource with a free keyword may be also indexing it according to a scientific thesaurus without even knowing it, while another user browsing a folksonomy will be able to find what he looks for because the engine can guide him through advanced relations.

b) **Services and tools** for the community

Every partner in a network - and partner could mean an institution, a working group or a single researcher - needs a specific share from the common information pool to work on it and will give back a specific contribution to the common pool. Services will need to become always more specific and tailored to specific needs.

A researcher / group / institution will set up their own working environment or dissemination platform by combining web services out of the community pool, and will in turn set up their own web services to expose the stream of the new information produced. All these web services will be interoperable and can be combined because of the use of common semantics.

At the most basic level, this might mean nothing more than using iGoogle for combining the personally needed streams, services and gadgets, and uploading one's own data to centralized repositories - such as those on the web "highstreet", like Google, Delicous, TheLibraryThing... - or harvesting data from the community. At an intermediate level, a common Open Source content management system will cater for all the needs of information retrieval and dissemination. At the sophisticated end there will be a network of specialized applications implementing web service clients, web service "endpoints" and different front ends. The old discussion about centralization and decentralization is finally obsolete.

The technological basis for semantic interoperability was laid when the W3C adopted RDF as the semantic description language. There have been gap years without really usable software to exploit this technology. This will quickly and dramatically change with Google now basing part of its information discovery on RDF snippets in text, and with one of the mainstream Open Source content management systems (Drupal) using RDF as the backbone for data management.

The CIARD framework of Application Profiles and subject vocabularies/ontologies will be of immense worth when these new tools start proliferating. Choosing tools that leverage standard vocabularies and RDF can allow for the implementation of really interoperable services. Therefore, the CIARD CMTF will create a community space in which suitable tools can be found, described, discussed and tested.

The final objective of adopting standards (vocabularies, protocols) and leveraging tools that enhance interoperability is to provide advanced, value-added, interoperable information services. In an RDF/semantic-web scenario like the one described, each service is a source for some services and a client for others. In order to orientate the community and support those who want to implement new services, the CIARD CMTF is designing the ARD RING (Routemap to Information Nodes and Gateways), a directory of information services in ARD with the specific objective of monitoring, describing and classifying the existing services, making them known and benchmarking them against interoperability criteria.

c) **People** – no future without information specialists – and **Institutions**.

The further development of technology and the onset of semantic interoperability do not mean that the Internet will become less complex. Always more processes and functions will be executed on the web. This also implies a constantly growing set of methodologies and technologies to use. Here are two examples from our area of expertise.

Example 1: publishing information.
Six years ago, someone who wanted to set up a very simple website, needed to know HTML and also needed some knowledge about how to put these HTML pages on a web server, not to speak about getting access to such a server. Today everyone without specific knowledge can setup a fairly nice blog on one of the existing "Web 2.0 services". Going further, he or she can buy some space on a hosting service and in two or three clicks install a highly sophisticated content management system.
The problem starts at this moment. How is she or he able not only to customize the system to best benefit him/her, but also for the maximum benefit of the community? No knowledge of server technology or programming languages is needed, but a high degree of knowledge about how that system works, how the web works and what has to be done to optimize this. This starts with a conceptional analysis of the service, the design of a semantic structure, the choice of semantic interoperability standards and includes but does not really limits itself to defining the RSS outputs of one's own system and the RSS feeds from other systems.

Example 2: retrieving information.
Do you need all the relevant information about a specific topic? Theoretically, everything is easier to find than 20 years ago; practically, you will always live with the doubt that you missed something. When I finished my PhD thesis in 1992 it was enough due diligence to have given the relevant keywords to the documentation services of the institutions and to have waded through the lists of references which they returned. If you missed something, that could be excused. Such excuses are not valid anymore. All information is on the Internet, isn't it? A lot of people develop their own search strategies and then stay with them despite changing environments. And these environments change quickly and steadily. Thomson Reuters is developing a business out of the dilemma. Thomson Reuters offers Web Intelligence: you ask, they answer, you pay. If you have information and want to mark it up, just deliver your content to Thomson Reuters, you will get it back with RDF metadata, free of charge. But Thomson Reuters keeps a copy of the metadata, accumulating a knowledge base to be used for future inquiries from clients. Any research institute would need to do something similar, but the people who would be able to do it are not there. Reference Librarians of the future will be different.

In these 2 examples only the easiest application scenarios are covered. There are others. Limiting ourselves only to our own environment (agricultural science and technology): how to set up a virtual working space between two research groups? How to use distributed computing power? How to link web services from partners to your own system?

These all are information management issues, not information technology issues. Information professionals, who are able to guide and assist researchers and policymakers in this area, are still very rare. We also see the future of libraries in research institutions in exactly this role. At the same time, the institutions in which these people work have to embrace these issues appropriately, by adequately resourcing their specialist IM units as well as by creating the enabling policies that provide the environment for such activities. These aspects are also areas of action for the CIARD initiative through the Advocacy Task Force.